機械学習アプローチに基づく生物学データ解析法に関する研究

AUTHOR(S):

烏山, 昌幸

CITATION:

烏山, 昌幸. 機械学習アプローチに基づく生物学データ解析法に関する研究. 京都大学化学研究所スーパーコンピュータシステム研究成果報告書 2014, 2013: 59-60

ISSUE DATE:

2014-03

URL:

http://hdl.handle.net/2433/186388

平成 25 年度 京都大学化学研究所 スーパーコンピュータシステム 利用報告書

機械学習アプローチに基づく生物学データ解析法に関する研究
A study on machine learning approach for biological data analysis

化学研究所バイオインフォマティクスセンター 生命知識工学研究領域　烏山 昌幸

背景と目的

タンパク質の相互作用ネットワークやシグナル伝達経路などバイオインフォマティクスにおいてはグラフとして表現されるデータが解析の対象となることは多い．一方で，遺伝子の発現量に代表されるような数値型のデータも生体内のメカニズムの理解のために重要な情報を含んでいることは広く認識されている．そのため，グラフと数値という異なるタイプのデータを統合的に解析する統計的な手法に対する需要が高まっている．

数値データを解析する統計的な基礎手法としては主成分分析や因子分析と呼ばれる多変量の共分散構造を解析する方法が知られている．これらの手法は古典的な手法ではあるが近年の大規模データ解析の場面においても汎用的な基礎手法として利用されている．ただし，これらの方法ではグラフの情報を含めることはできない．一方で，グラフを解析するためのアプローチとしてはグラフ理論に基づく方法が広く用いられている．スペクトラルクラスタリングに代表されるこのアプローチではグラフの構造を考慮しつつも，グラフのエッジに重みをもたせることで数値データの情報も扱うことができる．しかし，グラフ理論に基づくがゆえにデータの統計的な性質 (協調的な変動など) の解釈に関してこの方法で解析することは難しい．

検討内容

本研究ではグラフと数値データの情報を考慮した基礎解析手法の開発を目指し，グラフィカルガウシアンモデル (GGM) と呼ばれるモデルに基づく汎用アルゴリズムの設計を行った．

GGM においてはグラフの構造をガウス分布の条件付き依存関係として導入することができる．我々は GGM によって推定されたガウス分布が数値データのグラフ上での変動を表現していると解釈し，そこからグラフ上での相互作用と数値データの関係性を読み解くための基礎アルゴリズムの導出を行った．具体的には GGM によって推定されたガウス分布に対して因子分析を適用することで，グラフ上の共分散構造に対して解釈性の高い表現を獲得できることを示した．この手法を以降では Graph FA (Graph Factor Analysis) と呼ぶ事とする．

因子分析のモデルは以下のような線形式で表現される

$$x = Af + \epsilon,$$

ここで x は発現量等の数値データであり多変量のベクトルである，A は因子負荷行列と呼ばれる x の共分散構造を説明する行列，f は潜在的な因子を表現しており x より低い次元数を持つベクトルであり，ϵ は各次元が独立なベクトルである．通常の因子分析では与えられた数値ベクトル x の観測値集合を用いて推定を行う．特に A は共分散構造の要約として x の各次元の共変動に関する情報を持っている．Graph FA では GGM によって推定されたガウス分布からこの A を推定することでグラフの構造に沿った，つまりグラフ上で隣接する変数同士の共変動情報が観察できることを解析的に導出した．

結果

Graph FA がグラフ上の共分散構造を適切に抽出することを示すために実験による比較を行った．ここでは遺伝子発現量のデータとして乳がんに関する研究で用いられたものを使用する [1]．また，それらに含まれている遺伝子に対応するタンパク質相互作用のネットワークを Pathway Commons[2] から取得した．

ここではいくつかの手法を用いてタンパク質のネットワークから発現量データ上で大きな相互作用が観察される部分ネットワークを抽出するタスクを考える．Graph FA であれば A の各列ベクトルから値が大きい要素に対応するネットワークを抽出する．比較する他の手法でもほぼ同様の手順で部分ネットワークを抽出する．この場合，抽出されたネットワークが生体内で協調的に働いているような部分であることが望ましいと考えられる．このことを評価するためにここでは GO term enrichment analysis[3] と呼ばれる方法を用いた．GO term は遺伝子に対して生物学的な働きごとにアノテーションを付加したものであり，今回であれば抽出されたネットワーク上で隣り合った遺伝子が多くのアノテーションを共有していることが自然だと思われる．

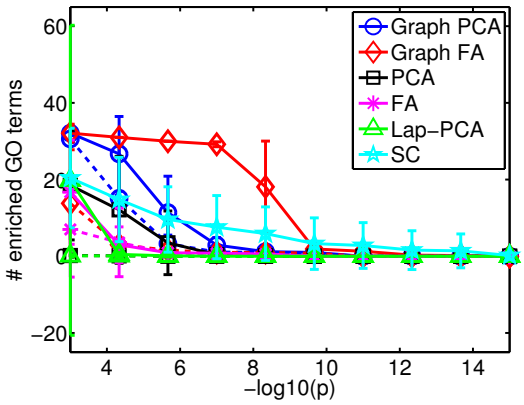

図 1: GO term enrichment analysis による比較．Graph FA, Graph PCA: 提案手法．FA と PCA はそれぞれ因子分析と主成分分析を表す．Graph PCA は因子分析の近似として主成分分析を用いたもの．PCA, FA: 主成分分析と因子分析．Lap PCA: スペクトラルグラフ理論を用いて正則化された主成分分析．SC: スペクトラルクラスタリング．

図 1 は実験結果の一部を示す．横軸は超幾何検定における p 値の対数をとった値であり，縦軸はその p 値で有意とされた GO term のうちグラフ上で隣接していたものの数である．Graph FA は他の手法と比較して多くの GO term を共有していることがわかる．

考察

今回はグラフと数値データの統合的解析を行う基礎手法として GGM と因子分析に基づく手法 (Graph FA) を開発した．実験的には Graph FA が抽出した部分ネットワークは GO term を多く共有しているという観点から生物学上でも何らかの意味を持つ遺伝子集合である可能性が示唆された．ただし，具体的にデータの内容 (今回の場合であれば乳がん患者の発現量) に関連した解釈はこれからの課題であり，このアプローチの正当性を実証する必要がある．

参考論文

[1] M. J. van de Vijver, and et al. A gene-expression signature as a predictor of survival breast cancer. *New England Journal of Medicine*, 347(25):1999–2009, 2002.

[2] E. G. Cerami, and et al. Pathway Commons, a web resource for biological pathway data. *Nucleic Acids Research*, 39 (Database-Issue): 685–690, 2011.

[3] M. Ashburner. Gene ontology: Tool for the unification of biology. *Nature Genetics*, 25: 25–29, 2000.